REPRINTED FROM
APRIL 2008
VOL. 22 | NO. 4
WWW.THE-SCIENTIST.COM

TheScientist
MAGAZINE OF THE LIFE SCIENCES

THE NCI'S BIOINFORMATICS NETWORK, caBIG, INTEGRATES CANCER DATA FROM ACROSS THE UNITED STATES.

ITS GOAL: TO SPEED THE TRANSITION FROM RESEARCH TO THERAPY

Heading for the BIG Time

By Kenneth Buetow
Artwork by Brendan Monroe

I was at the National Cancer Institute's (NCI) Intramural Program Scientific Retreat this past January listening to a plenary presentation by Cambridge University's Bruce Ponder, when a fascinating question caught my attention. Ponder described how variation in a gene called fibroblast growth factor receptor 2 (FGFR2) is associated with breast cancer. Judah Folkman, a legend among cancer researchers, stood up with a question for Ponder: Has anyone looked at the role of endostatin in breast cancer susceptibility? Endostatin is part of the same network as FGFR2, he explained; moreover, endostatin is located on chromosome 21, and trisomy 21 is protective against breast cancer. Ponder replied that his group had never examined the endostatin locus.

As quickly as the session ended, I stepped up to a Web browser and connected to the online resource caBIG (Cancer Biomedical Informatics Grid). By simply entering the endostatin locus, I was able to see that Folkman's scientific ▶

TheScientist
MAGAZINE OF THE LIFE SCIENCES

hunch was right on target: Multiple variants within the locus are significantly associated with breast cancer, and those loci are protective. Sadly, Judah Folkman passed away two days later, before I had a chance to share with him the product of his insight.

caBIG is a response to a desperate need. From my position as a senior cancer researcher at the NCI, groundbreaking observations and insights in biomedicine are accumulating at a dizzying rate. However, from the perspective of the approximately 1.4 million US patients who will hear their physicians say, "You have cancer," progress is unacceptably slow. Something needed to be done to expedite the transformation of scientific findings into clinical solutions.

Four years ago, I and my colleagues at the National Cancer Institute responded with the launch of caBIG (http://caBIG.cancer.gov) – a smart, World Wide Web of cancer research. Through the collaborative effort of member cancer centers, we collectively created more than 40 tools to squeeze the most out of cancer data, and a new, international infrastructure to connect the data. Researchers can use sophisticated tools to query multiple databases of raw data in order to generate or validate hypotheses. Already, researchers have published more than 45 articles, and we expect those numbers to grow as the value of these tools becomes obvious.

caBIG IS A RESPONSE TO A DESPERATE NEED.

Biomedical researchers struggle to meaningfully integrate their findings. Cancer is an immensely complex disease and in order to get a sense of the big picture, scientists need to combine observations from genomics, proteomics, pathology, imaging, and clinical trials. There was, however, no systematic way to do this. Encouraged by the support of our community and spurred to the challenge by our advisory boards, we set out to put a new set of tools into the hands of scientists – tools that would allow them to manage and understand the tsunami of biomedical data becoming available.

The caBIG was conceived in 2003 and born in the spring of 2004. It is indeed a big idea: to develop a state-of-the-art informatics platform that provides researchers all the capabilities they'd need to fight the "war on cancer." A large-scale, global concept for connectivity such as caBIG was unheard of in biomedicine in 2004 and is still foreign in most research domains today.

So, how did we develop caBIG? Given the urgency – more than 500,000 cancer deaths occur annually – we needed to start fast and learn quickly. My team at the NCI organized and launched a developmental "pilot phase" with NCI Cancer Centers, a collection of more than 60 long-standing research communities distributed throughout the country. These Centers had a limited history of cross-center collaboration. Could they work together? How many would be willing to be pioneers?

To answer this we went on the road to find out what cancer centers needed and what they were willing to share. The response was overwhelming. It became immediately clear that we needed to create a dynamic process that allowed rapid adjustment and growth. Our approach was to create a cross-institutional, virtual community composed of "workspaces." These workspaces would focus on topics ranging from creating a virtual tissue repository, to building tools that incorporate different data sources in research. Individuals, organizations, and institutions would work together to contribute applications, infrastructure, data, and

BRENDAN MONROE / WWW.BRENDANMONROE.COM

insights. Participants could benefit directly, from the collective expertise of this international collaboration.

For this virtual community to succeed it was important to embrace the individual diversity of members and to connect them, as opposed to creating one big central resource where everyone needed to place their information. As such, caBIG focused on providing tools and infrastructure that could be run by individual laboratories, organizations, or institutions and connect electronically through the Internet. This strategy is called standards-based interoperability, and caBIG has realized it through a services-oriented architecture called caGrid. It is worth noting that caBIG adopted international standards where they existed and extended them as needed to address new problems.

We used all possible mechanisms to invite, engage, and sustain relationships among participants. Weekly teleconferences, supported by Webcasts of presentations and countless listsrvs kept members of the community connected; these discussions are now archived at http://caBIG.nci.nih.gov. Today, these ongoing, virtual interactions continue, augmented by regular face-to-face meetings and a weekly e-newsletter, called "What's BIG this Week," which distributes meeting schedules and key discussion topics for the upcoming week.

The workspaces attract a wide diversity of participants, including informatics experts, clinicians, bench researchers, patient advocates, and senior executives at pharmaceutical companies. More than 190 organizations have participated in caBIG, including NCI-designated Cancer Centers, federal agencies, academic institutions, not-for-profit organizations, and biotech and pharma companies. Essentially, a corps of more than 1,000 individuals is finding creative ways to use caBIG and to sharpen its tools. The initiative isn't meant to serve only the "big science" centers. The tools empower individual laboratories, organizations, and institutions to innovate through traditional (often single investigator-driven) research programs.

Andrea Califano of Columbia University, for example, developed a software program called cancer Workbench (caWorkbench). The program allows an investigator to electronically grab and analyze microarray data from different sources and perform multiscale analysis of genomic and cellular networks. It has become one of the tools that can be freely downloaded or accessed through the caBIG Web browsers.

We intend for researchers to share not only their software, but also their data, where possible. There are obvious challenges associated with data sharing ▸

A sampling of how you can use caBIG:

BIOBANKING MANAGEMENT WITH caTISSUE

- Access a library of well characterized, clinically annotated biospecimens.
- Use the tool to keep an inventory of a user's own samples.
- Track the storage, distribution, and quality assurance of specimens.

MOLECULAR DISCOVERY

- The caIntegrator – Combines proteomics, gene expression, and other basic research data with clinical trial results.
- The caArray – A repository of microarray data, which allows users to submit and annotate their data.
- The geWorkbench – Allows integration of gene mircoarray data from multiple manufacturers and permits analysis and visualization of that data.

CLINICAL TRIAL SOFTWARE BUNDLE

- Track clinical trial registrations.
- Facilitate automatic capture of clinical laboratory data.
- Manage reports describing adverse events during clinical trials.

IMAGING ANALYSIS

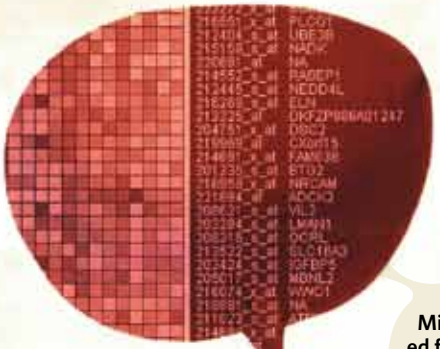
- Includes The National Cancer Imaging Archive, a repository for medical images including CAT scans and MRIs.
- Quantitative assessment of drug response with the archive's associated software.
- Images can be visualized with a tool that lets users annotate areas of interest.

If you need help getting started, you can find tutorials at http://caBIG.cancer.gov, or you can reach technical support at 888-478-4423.

caBIG IN ACTION

Here is a hypothetical case study that uses caBIG tools and databases:

In the cancer glioblastoma multiforma (GBM), does expression of key genes predict disease outcome? How might these genes interact in biological networks that would influence therapeutic targeting?

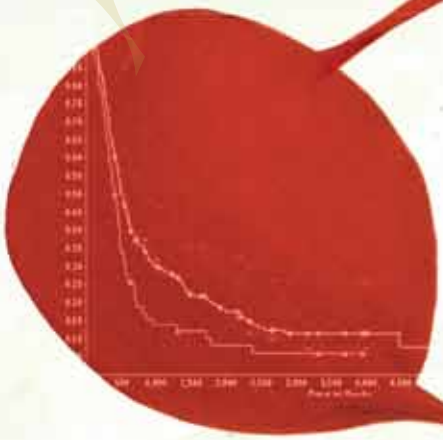


1: Data Source
Microarray data collected from patient samples in California, Texas and Tennessee are annotated with study details using caArray.

2: Data Source
Clinical outcomes from patients are managed and tracked through the clinical trial software bundle. Shown here is a post-surgical image shared through the National Cancer Image Archive (NCIA).

3: Data Integration Tool
The caIntegrator – which can be used online, or downloaded for internal use – combines and normalizes the various sources of data.

4: Find Gene of Interest
Portals leveraging caIntegrator can produce various reports, statistical analyses and graphs of the combined data sources. This graph shows differential survival of patients with GBM associated with the expression of the gene KITLG. Patients with low levels of KITLG have significantly longer survival than those with high levels.



5: Find Related Genes
Using the caBIG-compatible tool GenePattern tool served by a caBIG portal, it is possible to find other genes that have related patterns of expression in tumors. Shown here is an expression map of 45 additional genes that have expression patterns related to KITLG in GBM.

6: Protein Pathway Mapping
With a target set of genes in hand, researchers can use the Pathway Interaction Database to identify the complex network that underpins disease. This database provides annotated information on the connections between cellular pathways. In this example, the genes whose expression is correlated with KITLG are shown to form a complex network of interactions. Understanding these networks will be important to developing combinatorial strategies for targeted interventions.

between industry and academia, and between academic researchers vying for the same funding pools. Also, appropriate protections need to be provided for research participants who have generously donated information and material. One caBIG workspace, the data sharing and intellectual capital workspace, focuses on these issues. It integrates expertise from technology transfer specialists, legal counsel, ethicists, security experts, institutional review boards, privacy authorities, and the advocacy community, among others, to create frameworks that guide data sharing and address security and protection of human subjects.

At the end of our pilot phase, we had assembled a vibrant community, a rich collection of tools, and a unique infrastructure to connect and share. The next challenge is to see whether and how the broader universe of biologists and clinicians will adopt them.

One of the first research programs to use caBIG on a large scale was a project aimed at finding cures for brain cancer. Brain cancer is relatively rare in the general population but is the leading cause of cancer mortality in children. The median survival of patients is approximately one year, and the few long-term survivors face significant lifelong neurocognitive deficits. Arguably, survival has not significantly improved in more than a decade, with rarity representing one of the main barriers to progress – no individual investigator or institution sees enough cases to conduct clinical research.

To address this issue, the Glioma Molecular Diagnostic Initiative (GMDI) was launched. Led by Howard Fine of the NCI, multiple investigators from around the world have collaborated to more accurately characterize gliomas, the most common form of brain tumor, using immunohistochemistry, genetics, and molecular biology.

In partnership with the caBIG community, Fine's group created the Repository of Molecular Brain Tumor Data (REMBRANDT, see http://rembrandt.nci.nih.gov). By customizing the caBIG infrastructure, REMBRANDT provides an unprecedented opportunity to conduct in silico research, both for hypothesis generation and external validation. It can be used for gene discovery, elucidation of the role of pathways, and molecular target identification and validation. Use of the data is free and without expectation of coauthorship, coinventorship, and any other type of remuneration.

REMBRANDT has already paid dividends. Sun and colleagues have discovered that stem cell factor (SCF) is a critical angiogenic factor in the pathogenesis of malignant gliomas (*Cancer Cell* 9:287–300, 2006). The authors were able to correlate SCF to clinical outcome. This association was made painstakingly through extensive in vitro and in vivo studies. Today, that same relationship can be revealed by simply querying the large patient population in REMBRANDT. The results of the REMBRANDT analysis show that questions concerning gene expression and survival are accessible through a few intuitive clicks of a mouse. Any researcher can use the underlying caBIG software (caIntegrator DataMart) that powers REMBRANDT to integrate large databases of disparate data sources and to query them with a number of statistical tools (see infographic, p. 64).

Other tools are already generating novel scientific findings. Louise Showe's group at the Wistar Institute in Philadelphia was looking for a way to distinguish two cancers that had very similar histology, but which required very different treatment protocols. Her group wanted to distinguish the two cancers on a genetic basis. In order to perform her analysis, Showe had to integrate gene-expression arrays from four different institutions to get the volume of data she needed. In addition, the data to be integrated came from two kinds of Affymetrix chips. Showe and colleagues fine-tuned a tool called distance-weighted discrimination (DWD), which statistically manipulated the data from the different microarray platforms so that they could be analyzed as a single source. The group found a ▶

TheScientist
MAGAZINE OF THE LIFE SCIENCES

"The power of this project

IS THAT IT INTEGRATES MASSIVE, HETEROGENEOUS DATASETS USING A USER-FRIENDLY INTERFACE."

panel of 10 genes that positively distinguished the two cancers. The DWD tool is also now accessible to researchers on the caBIG Web site.

The caBIG community is also generating tools and infrastructure to connect collaborative networks and to make large databases easier to use. The NCI and the National Human Genome Research Institute are currently creating a comprehensive catalog of the gene changes that underlie multiple forms of cancer. Called the Cancer Genome Atlas (TCGA), it is a multidimensional, molecular characterization of cancer. In the pilot phase of the program, data on large scale sequencing, gene expression, DNA fragment copy number changes, loss of heterozygosity, and the epigenetic state of the genome are being generated on a common collection of biospecimens that will be annotated with rich clinical information.

The caBIG connects and integrates the terabytes of TCGA data generated at 11 different sites distributed throughout the United States. Using the caBIG infrastructure, data are shared electronically with the public. Out of concern for human subjects' privacy and protection, not everything is made public. A data-access committee reviews users and permits access to protected data only to those who qualify. However, with authorization, it is possible to use caBIG tools to comprehensively interrogate this unique data collection.

The power of this project is that it integrates massive, heterogeneous datasets using a user-friendly interface. For example, a researcher can browse the mutations found in a particular cancer using the Cancer Genome Workbench (http://cgwb.nci.nih.gov) and then use the Pathway Interaction Database (http://pid.nci.nih.gov) to look at how those mutations come together in a cellular pathway (see infographic, p. 64). Using the TCGA portal, the researcher can combine the mutations with data on gene deletions and amplifications to study how they interact in multiple pathways.

Further, using tools similar to those in REMBRANDT, researchers can examine the joint effects of mutation and gene expression in altering survival time. By simply selecting from a menu of gene mutations observed in the sample set and choosing any genes characterized in the genome-wide expression studies, a user can see how different components interact to alter disease outcome.

While the current sample sizes in the TCGA are too small to generate anything but intriguing hypotheses, it is already generating provocative observations. For example, SCF expression appears to alter prognosis in glioblastoma multiform (as above), but only in individuals who have a *p53* mutation.

One goal we have here at NCI is to connect, by the year 2010, all NCI comprehensive and community cancer centers in the United States. The data collected at each center will be shared (as appropriate), and all multicenter clinical cancer trials will connect to each other electronically and to the Food and Drug Administration. Perhaps most ambitiously, institutions will be collaborating and publishing studies through caBIG. The longer-term goal is to extend the standards, infrastructure, and vocabularies that were pioneered for cancer to connect the entire biomedical enterprise, regardless of the disease studied.

As a result, I envision huge advances that will bring a new era in molecular medicine. Benefits include early identification of disease-causing genes, which will enable us to delay or prevent the progression to clinical symptoms. Subgrouping of diseases by genetic biomarkers will allow us to predict how a disease will advance and how amenable it will be to therapeutic options. Moreover, the capacity to monitor patient response to treatment will obviate useless approaches and make it possible to prescribe "the right drug for the right patient at the right time."

It's a daunting challenge. But we need – and intend – to move at warp speed to serve the patient community. By putting the right data in the right hands at the right time, we can quicken discovery, eliminate unnecessary redundancy of research, and better understand clinical success and failure. This comprehensive, integrated view fully embraces the complexity of cancer and will allow us to determine rationally how to combine multiple interventions. ■

Have a comment? E-mail us at mail@the-scientist.com

Kenneth Buetow is the NCI Associate Director for Bioinformatics and Information Technology, a laboratory chief in the Laboratory of Population Genetics at the National Cancer Institute, and he is the founder of the caBIG project.